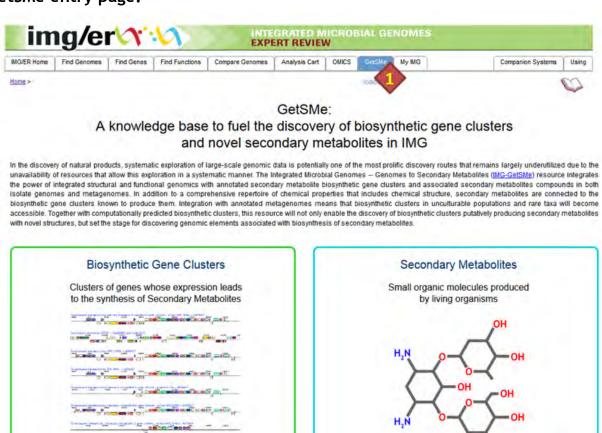
Introduction to IMG GetSMe

Genomes to Secondary Metabolites (GetSMe) is the IMG resource dedicated to the analysis and discovery of biosynthetic gene clusters (BCs) and associated secondary metabolites (SMs) in the genomes available in the IMG database. This doscument describes the different features of GetSMe, as well as some workflow examples to assist the user in taking full advantage of the tools and data available in IMG.

GetSMe entry page:



Though the entry page, which is accessible through the GetSMe tab in IMG (1), a user can navigate to four user interfaces that serve different purposes:

907,144 predicted

Search BCs

• Browse BCs (2): This leads to a page where the user can get summaru statistics regarding the different attributes of BCs in GetSMe, such as BC length, PFAM content, BC type etc.

1,108 structures

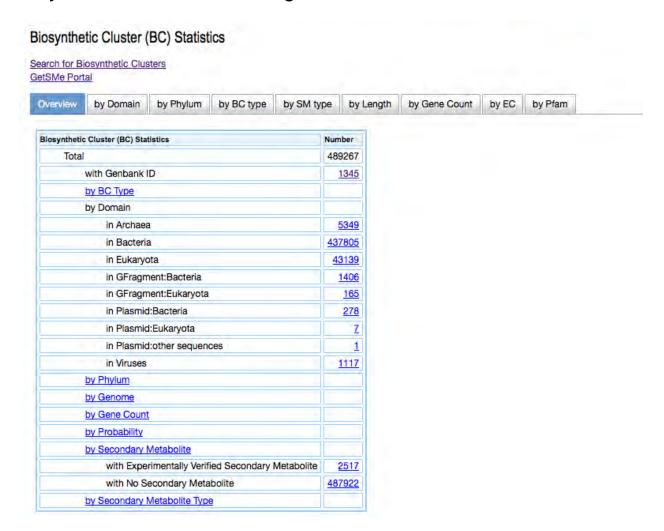
earch SMs

• Search BCs (3): The user can use this interface to search all BCs based on several text and numeric attributes.

- Browse SMs (4): Similarly to (2), the user can view summary statistics for the chemical compounds designated as secondary metabolites (SMs) through their association with BCs.
- Search SMs (5): In addition to the functionality of the Search BCs (3) feature, Search SMs allows for queries to be made using chemical structures (in the form of a SMILES string).

All these functions are also available through a drop-down menu that appears when the user hovers their cursor over the GetSMe tab (1).

Biosynthetic Clusters Statistics Page



The entry page to the BC portion of IMG GetSMe is partitioned in 9 tabs based on different attributes. These tabs are:

- 1. **Overview.** This is the default tab contains a table giving a high-level snapshot of the data available in IMG GetSMe.
- by Domain. The number of BCs in IMG GetSMe is grouped by domain and presented in a bar table. Each bar is clickable and links to a tabular listing of the BCs within the chosen domain.

- 3. **by Phylum.** Similar to the "by Domain" tab, but summarizes data based on Phyla, instead.
- 4. **by BC type.** All predicted BCs were acquired through the implementation of the ClusterFinder algorithm¹. All the predicted BCs were then fed to the antiSMASH tool² which assigns a type, when possible, to the BC based on its enzymatic composition. These annotations are available in IMG GetSMe through the "BC type" label. Whenever more than one types were assigned to a BC, the types were sorted and concatanated into a semi-colon separared string, e.g. "bacteriocin;lantipeptide;t1pks;thiopeptide". These annotations are summarized in the "by BC type" along with the number of BCs in each predicted category.
- 5. **by SM type.** This tab pertains only to BCs retrieved from the GenBank database and that are connected to at least one compound with a known chemical structure. The SM type refers to the MeSH Classification associated with the chemical structure in the PubChem Compound database.
- 6. **by Length.** Histogram grouping BCs based on their length in bases.
- 7. **by Gene Count.** BCs are grouped in this graph based on the number of genes they contain. Two graphs are presented, one for experimental and one for predicted BCs.
- 8. **by EC.** This tab contains an expandable tree which groups BCs based on the enzymatic functions of the gene products in the BC. The enzymatic function is annotated as an Enzyme Commission (EC) number, which follows a hierarchical structure of classifications. The number next to the EC number referes to the number of distinct BCs that contain at least one gene whose product is annotated with that enzymatic function.
- 9. **by Pfam.** Data are summarized by the protein domain classifications (Pfam numbers) that each BCs contains. This tab contains both a table and a pie chart summarizing the occurrences of Pfam numbers. These summary data are also partitioned based on whether they are associated with experimental or predicted BCs.

¹ Cimermancic, Peter, et al. "Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters." *Cell* 158.2 (2014): 412-421.

² Blin, Kai, et al. "antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers." *Nucleic acids research* (2013): gkt449.